Clustering and Analysis of Tweets Related to Petrobras

Demetrius M. Murato[1], Bruno S. dos Santos[1], Rafael Henrique Palma Lima[1]

[1]Departamento Acadêmico de Engenharia de Produção – Universidade Tecnológica Federal do Paraná (UTFPR)
Caixa Postal 86036-370 – Londrina – PR – Brazil

demetriusmurato@alunos.utfpr.edu.br, brunosantos@utfpr.edu.br, rafaelhlima@utfpr.edu.br

Abstract. *This study aimed to cluster and analyze tweets associated with Petrobras, exploring its meaning and user profiles on social media to understand their impact on financial markets. The research applied a workflow including the data collection from Twitter's API (current X), preprocessing of tweets using Python libraries, word vectorization via Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), Principal Component Analysis (PCA) to reduce matrix dimensionality, and the K-means clustering technique. A total of 840 preprocessed tweets were clustered and analyzed for patterns related to Petrobras. Five clusters were identified in the initial analysis with no dimensionality reduction, showcasing differing characteristics, while the subsequent PCA-based analysis yielded three clusters showing contrasting themes in tweets. The PCA-based analysis showed grouped tweets about the market and economy (cluster 0), while cluster 1 was related to political concerns. Limitations included reliance on publicly available Twitter data, constraints due to the quantity and nature of tweets, and potential biases in sentiment analysis due to informal language and sarcasm. The research underscores the potential of unsupervised machine learning techniques in analyzing sentiments and user profiles related to financial markets. Insights derived from tweet clustering could aid investors in gauging market sentiment.*

Resumo. *Este estudo teve como objetivo agrupar e analisar tweets associados à Petrobras, explorando seu significado e perfis de usuários nas redes sociais para compreender seu impacto nos mercados financeiros. A pesquisa aplicou um fluxo de trabalho que inclui coleta de dados da API do Twitter (atual X), pré-processamento de tweets usando bibliotecas Python, vetorização de palavras via Bag-of-Words (BoW) e Term Frequency-Inverse Document Frequency (TF-IDF), Análise de Componentes Principais (PCA) para reduzir a dimensionalidade da matriz e a técnica de agrupamento K-means. Um total de 840 tweets pré-processados foram agrupados e analisados em busca de padrões relacionados à Petrobras. Cinco clusters foram identificados na análise inicial sem redução de dimensionalidade, apresentando características diferentes, enquanto a análise subsequente baseada em PCA obteve três clusters mostrando temas contrastantes em tweets. A análise*

Cadernos do IME - Série Informática
e-ISSN: 2317-2193 (online)
DOI: 10.12957/cadinf.2024.82401

baseada no PCA mostrou tweets agrupados sobre o mercado e a economia (cluster 0), enquanto o cluster 1 estava relacionado com preocupações políticas. As limitações incluíram a dependência de dados do Twitter disponíveis publicamente, restrições devido à quantidade e natureza dos tweets e potenciais desvios na análise de sentimentos devido à linguagem informal e ao sarcasmo. A pesquisa destaca o potencial das técnicas de aprendizado de máquina não supervisionado na análise de sentimentos e perfis de usuários relacionados aos mercados financeiros. As informações derivadas do agrupamento de tweets podem ajudar os investidores a avaliarem o sentimento do mercado.

Keywords: Text mining, Clustering, Petrobras, Principal Component Analysis.

1. Introduction

The Brazilian stock exchange, Brasil, Bolsa e Balcão (B3), currently holds approximately 3.3 million accounts [B3, 2020], which traded a total of R$6.45 trillion in 2020 through financial asset negotiations [Money Times, 2021]. Among the assets traded, stocks stand out due to their liquidity and earnings potential. Investors use both fundamental analysis, which looks for the intrinsic value of companies, and technical analysis, which is based on graphic patterns for short-term decisions [da Silva et al., 2020].

Specifically in short-term decisions, machine learning (ML) techniques have recently gained prominence in stock market predictions, showing significant growth in global research [Kumbure et al., 2022; Henrique et al., 2019]. ML, as a subfield of artificial intelligence, improves predictive models by analyzing structured quantitative data or unstructured data such as texts [Raschka, 2015].

In the context of stock market predictions, structured data is often transformed into inputs to identify trends or predict market movements [Adegboye and Kampouridis, 2021]. Furthermore, sentiment analysis using unstructured data such as social media posts has become important for assessing investor sentiment and its impact on stock prices [Yadav et al., 2019; Seong and Nam, 2021], and the application of classification or clustering tasks helps researchers and analysts to interpret unstructured data from social media and online news [Khadjeh Nassirtoussi et al., 2014]. When applied in a document context, the unsupervised techniques provide a logical and understandable framework for organization, navigation, and search [Braga, 2018], extending to social media where each post can be considered a document. Despite the differences between financial analysis and natural language processing, both are applicable in unsupervised learning [Chen et al., 2021].

Numerous studies utilize ML techniques in the 'stock market' domain. For instance, Nam and Seong (2019) proposed a prediction method employing analysis of influential causes in Korean stock market-related news. Nizer and Nievola (2012) applied text mining in news to predict asset volatility in the Brazilian stock market, while Carosia et al. (2020) conducted sentiment analysis based on Twitter data during

the 2018 electoral period. Moreover, Oliveira et al. (2017) utilized prediction techniques in returns, volatility, traded volume, and sentiment indices from two existing Twitter accounts. Shi et al. (2019) proposed a different approach, focusing on sentiment contagion analysis in investor forums, estimating contagion model parameters to analyze sentiment formation at an individual level. Sun et al. (2020) studied over 20,000 tweets from the Chinese social media and microblogging platform Sina Weibo, examining user profiles in two distinct clusters and their influence on the stock market. Katayama and Tsuda (2020) applied sentiment evaluation in Japanese news to apply acquired knowledge in investment strategies for individual stocks.

Despite established studies in data related to market analysis, whether company-based or stock-specific, none have used clustering techniques on actions or the Petrobras-related market specifically with data extracted solely from the Twitter platform. The closest articles to the aim of this research are the applications by Lima et al. (2016) and Akita and da Silva (2023), which used Petrobras-related tweets to classify sentiments, but not to cluster similar tweets. The work proposed by de Oliveira et al. (2013) employed neural networks to enhance predictions of preferred share PETR4, albeit with structured data.

Thus, this article aims to describe and analyze data related to the publicly traded company Petróleo Brasileiro S.A., also known as Petrobras, by grouping posts made on the social media platform Twitter (current X), focusing on preprocessing to analyze quantitative and qualitative characteristics. The research motivation and data collection period (September 21st and 22nd, 2021) stemmed from significant stock price volatility and drops in the preceding days. Additionally, global market rumors indicating a potential oil barrel price drop, confirmed in news on the 20th, and turmoil within Brazil's Transpetro—a Petrobras subsidiary focused on oil and derivative transportation—resulted in its president's resignation the next day. These two news pieces were reflected in Twitter discussions.

2. Materials and Methods

The data in this study, regarding the company Petrobras and terms directly related to it, were collected from the social media platform Twitter (current X) through connection with Twitter API v1.1, a platform designed for developers [Twitter Inc., n.d.].

2.1. Twitter (X)

With significant similarity to data obtained from news platforms, social media data can manifest in different forms, with Twitter being the primary platform for accessing these datasets [Kumbure et al., 2022].

Given this and considering it as one of the largest social networks globally concerning active user numbers, confirms the impact and relevance this social network has on society, particularly in information dissemination. This holds when it comes to the financial market, with existing accounts of investors, public figures, or even management entities discussing the subject matter [Info Money, 2017].

Thus, understanding the influence Twitter wields in society and the comments and opinions of individuals available on the platform, tweets related to the publicly traded company Petróleo Brasileiro S.A. were collected. An important point to note is that these data, being sourced from a social network, are not standardized and often

exist in informal language, laden with abbreviations due to the 280-character limit per tweet. Additionally, emojis and emoticons (messages in image or icon formats) might be present within these texts [Twitter Inc., n.d.].

Specifically, Twitter functions as a sort of microblogging platform where users post messages that may contain hashtags to highlight a topic for discussion [Ahmed et al., 2017], photos, videos, and even links to other websites. It also allows users to interact with each other through responses, sharing, 'likes,' retweets, and mentions of other users [Twitter Inc., n.d.].

2.2. Twitter API

Tweets are visible and searchable by anyone worldwide, even if the user is solely browsing through message updates. Twitter collects user's personal information, such as the type of device being used and the Internet Protocol (IP) address, and additional information like location.

The Twitter API enables developers to access thousands of past tweets and even real-time data, specific user profiles, geographic trends, and word searches, among other tools. These resources can be accessed using the TwitterSearch library, allowing retrieval of text and numerical data for Python in the '.json' extension (JavaScript Object Notation).

It is important to note that there are different versions of the Twitter API, differing in the number of tweets that can be extracted within a period, additional tools, and even paid versions. This study utilized version 1.1, which is free, offering an academic limit of 10 million tweets per month for each project or 500,000 for standard projects, with variable update request rates [Twitter Inc., n.d.]. In summary, Twitter API provides open access for researchers and fits within the scope of sentiment analysis due to its consistency in post-character length [Sarram and Ivey, 2022].

2.3. Description of the phases and the methods

For the experiments, Petrobras was chosen as the subject of study due to its frequent mention in financial market news. This arises from several characteristics exhibited by the company, such as its worldwide recognition, high liquidity, reaching an average daily trading volume close to R$2 billion in 2021, making it the second most traded company on B3 in 2022, concerning only preferred shares [Info Money, 2022]. The company also has a history of resilience despite past setbacks, like the change in its dividend-cut policy in 2012 and 2013, resulting in a significant drop in ordinary and preferred shares (PETR3 and PETR4, respectively) and a subsequent reduction in its market value [Jesus Júnior et al., 2017].

Initially, a filter was applied for subjects related to the company through the 'keyword' parameter, i.e., messages containing predefined words ('Petrobras', 'petr4', 'petr3' and 'petróleo') or related to them, via retweets or replies to the main tweet. Consequently, the resources provided by the TwitterSearch library were used, allowing the collection of 3,598 tweets conducted between September 21st and 22nd, 2021—chosen due to the significant volatility in stock prices during the period, along with events like the oil barrel price drop on the 20th and the Transpetro president's resignation on the 22nd, a Petrobras-related company focusing on oil and derivative transportation logistics. It's essential to note that apart from setting a search parameter,

defining the sought characteristics was necessary. In this case, it involved information on the date and time of creation, user identification, tweet content, number of 'likes' received, and finally, the count of retweets.

As the data was directly extracted from Twitter using the Twitter API, the set did not come in a suitable format for analysis, requiring preprocessing since social media texts are typically informal and non-standardized. With the assistance of Python's Json and Pandas libraries, the original data was transformed into a dataset comprising 3,598 posts. The subsequent step involved preprocessing the texts of each tweet, starting with the removal of duplicate tweets, conversion to lowercase, elimination of links, unwanted characters, and punctuation, utilizing resources from the Regular expression (Re) and Beautiful Soup (BS4) libraries. After preprocessing the database, 840 tweets remained due to duplicate text removal, enabling the Spacy, Unidecode, and Natural Language Toolkit (NLTK) libraries to remove stopwords (a list of undesirable or irrelevant words for analysis), eliminate accents, and transform words into their respective roots by suffix elimination.

Subsequently, the resources of the Scikit-Learn library were used to apply the Bag-of-Words (BoW) technique and define the weights of each word through Term Frequency-Inverse Document Frequency (TF-IDF). The first method maps the document into a vector space and quantifies the number of occurrences of a term, while the second associates the frequency of a term's appearance in a document and the fraction of documents containing it [Kononova et al., 2021].

In the study, applying TF-IDF to the 840 posts resulted in a weight matrix with 840 rows (tweets) and 2,355 columns. This transformation constructs variables from each unique token (word) considered within the document, resulting in high data dimensionality. To improve this situation, further data cleaning was conducted through the application of two functions, described as follows:

1) Elimination of words that appeared fewer than three times in the entire dataset (after applying BoW), totaling an exclusion of 509 words and resulting in 1,846 tokens;
2) Removal of tokens that were strings of numbers not excluded during preprocessing, due to being alongside letters without any space and not identified in the initial data treatment. This resulted in a weight matrix with 840 rows and 612 columns.

The clustering technique was applied to two distinct datasets to allow a comparison:

- Dataset with all tokens resulting from preprocessing (i.e., after TF-IDF application), resulting in a matrix of 840 instances and 612 tokens;
- Reduced dataset using Principal Component Analysis (PCA) as the high dimensionality obtained from tokenization could affect the clustering algorithm's accuracy, generating noise [Mohamed, 2020].

2.4. First clustering approach

Before the analysis of the first clustering, the Within Cluster Sum of Squares (WCSS) metric was evaluated in the weight matrix by applying the K-means algorithm. Raschka

(2015) describes this method as a step in the K-means algorithm that aims to minimize the sum of squared errors within the cluster and allows for a graphical visualization of data behavior with different numbers of clusters. Another evaluated metric was the Silhouette Score, which, like WCSS, also offers a graphical visualization of how well the data has been grouped [Naghizadeh and Metaxas, 2020]. Consequently, this enabled the comparison of results and the determination of the best number of clusters, identified by the parameter '*k*'.

Although there are other metrics for evaluating the number or validation of clusters, Al-jabery et al. (2020) cite that the Silhouette Score is one of the most common metrics for this purpose, and Arbelaitz et al. (2013) affirm its better results in several cases. On the other hand, WCSS is integrated into K-means and does not require any additional calculation or resource, making it quite popular and the default metric in various software [Naghizadeh and Metaxas, 2020]. Therefore, these two metrics were chosen for the present study.

The K-means algorithm belongs to the class of methods based on unsupervised learning, where there is no specific target class or variable, as seen in predictive models using supervised tasks [James et al., 2023]. However, it generally requires low computational effort to separate data based on their characteristics [Ulfenborg et al., 2021]. The input parameters for data separation were the weight matrix (840 x 612) and the value of *k*, resulting in the assignment of each tweet to a cluster.

Based on the previous steps, columns were created in the dataset containing the sum of weights for each tweet, the cluster assigned to the tweet, the average weight, and the word count analyzed for each tweet. For result analysis, a table was constructed with the most frequent tokens in each cluster.

2.5. Second clustering approach

The second clustering employed Principal Component Analysis (PCA), which resized the original dataset to fewer columns referred to as components. This column reduction is achieved by rotating orthogonal axes, aligning two or three axes in an ellipsoid shape, representing the direction of maximum variance of the studied phenomenon [Pompella and Dicanio, 2017]. The objective is to maintain maximum information in a smaller-dimensional matrix. In large datasets, the loss of information often does not hinder the model or the studied output [Singh and Srivastava, 2017].

Following the matrix transformation by PCA, the WCSS and Silhouette Score analyses were again applied, but this time on the new weight matrix. The K-means algorithm received, as an input parameter, the newly transformed weight matrix.

Subsequently, new columns were added to it, corresponding to:

- The sum of weights for each tweet;
- The cluster assigned to the tweet;
- The average weight of each instance;
- The word count analyzed for each tweet.

A word (token) table was created for each analyzed cluster. Finally, it became possible to compare the results obtained from the two datasets, i.e., clustering performed with and without initial resizing of the weight matrix obtained in TF-IDF.

The Figure 1 shows the workflow of the research.

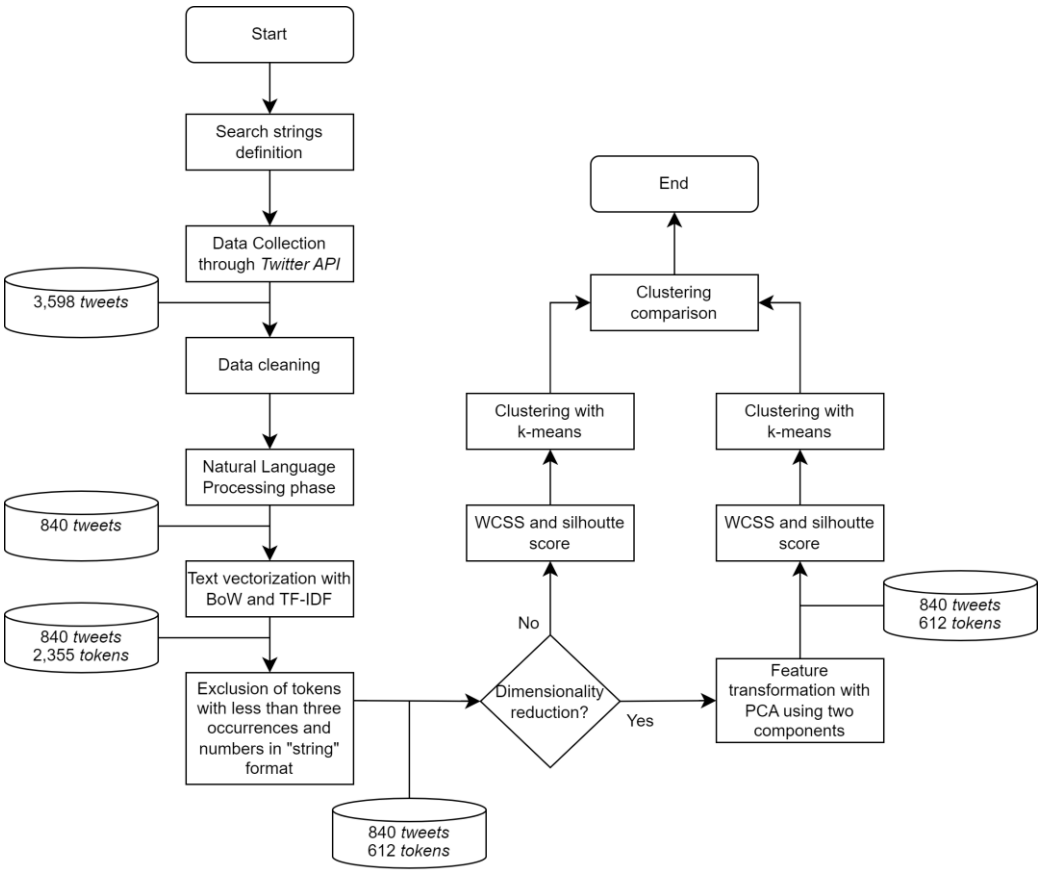

Figure 1. The research workflow

3. Analysis and Discussion of the Results

In this section, the results obtained from the proposed workflow depicted in Figure 1 are presented, encompassing the preprocessing steps through to the comparison and discussion of the achieved outcomes.

3.1. Preprocessing

As mentioned earlier, texts sourced from social media platforms are typically non-standardized. For instance, the term “Petróleo” can exhibit variations such as “PETRÓLEO”, “Petróleo”, “petróleo”, and “petroleo”, thereby expanding the dataset and consequently increasing computational effort. Hence, to begin normalization, duplicate tweets (retweets without new texts) were excluded initially, and subsequently, the data was converted to lowercase. Undesirable characters such as emojis, emoticons, punctuations, accents, and links were also removed. This was followed by the elimination of stopwords and suffixes, occurring in the final stage of preprocessing.

Table 1 illustrates examples of tweets before and after normalization.

Table 1. Tweets before and after normalization

Date	Original tweet	Number of favorites	Number of retweets	Normalized tweet
2021-09-22 14:40:22	Hoje, completamos 40 dias sem reajuste de preços da gasolina A nas refinarias da Petrobras e mesmo assim o preço aumentou… https://t.co/0AyxC9vqyy	70	10	hoj complet 40 dia reajust prec gasolin refin petrobr prec aument
2021-09-22 10:23:00	Lembra como o volume estava forte? PETR4 agora tá projetando só R$69,6 MM de ações ... 😕	25	5	lembr volum fort petr4 projet 69,6 acoe
2021-09-22 10:00:02	Os preços da gasolina praticados pela Petrobras hoje têm defasagem média de 6% em relação aos preços internacionais… https://t.co/ygcYh5g5X2	53	64	prec gasolin pratic petrobr hoj defasag medi 6 prec internacion

3.2. Weight matrix

The initial tokenization of preprocessed tweets was performed using BoW, resulting in a matrix with 840 instances (total tweets) and 2,355 columns, representing all unique terms (tokens) across all tweets, where n_ij denotes the number of times a term j appeared in a tweet i. Table 2 displays the top four terms that most frequently recurred in tweets according to BoW. It is important to note that a term can appear more than once within the same tweet.

Table 2. Most frequent terms in preprocessed tweets

Token	Frequency
petrobr	286
petrole	234
brasil	78
prec	64

The TF-IDF was applied to assess the significance of terms concerning the dataset, resulting in a weight matrix with the same dimensions as obtained by BoW, although cell values were calculated differently (refer to the "Materials and Methods" section).

The application of BoW and TF-IDF enabled the visualization of all tokens in columns, their occurrences, and weights, highlighting occasional appearances of numbers in a single tweet, such as the price of gasoline in a specific city, and numerous words that appeared infrequently in the dataset, both resulting in negligible weights. Consequently, two functions were employed to eliminate data with these characteristics.

The first function removed all tokens representing numbers, while the second function eliminated terms that occurred fewer than three times in the entire dataset, a value empirically defined. As a result, the weight matrix underwent a reduction of approximately 74% in the total number of terms, ending with 840 instances and 612 tokens.

3.3. Clustering with the original weight matrix

Before conducting the final clustering for interpretations, the WCSS (Within Cluster Sum of Squares) and Silhouette score metrics were employed to assist in determining the number of clusters. In this instance, a range of cluster numbers (parameter *k*) up to 10 was explored using the weight matrix (840 x 612) obtained after applying TF-IDF.

Figure 2 illustrates the WCSS performance, while Figure 3 represents the Silhouette score performance.

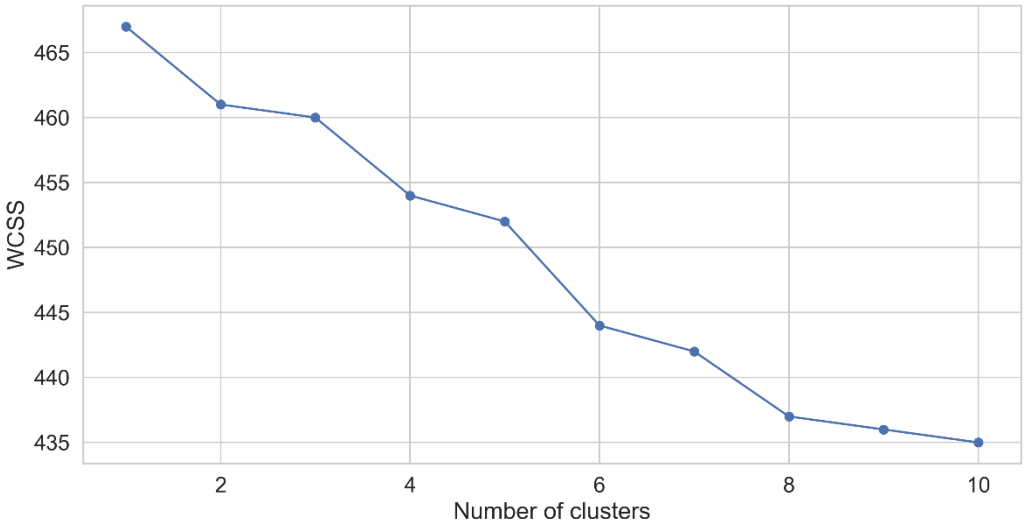

Figure 2. WCSS results in the original weight matrix

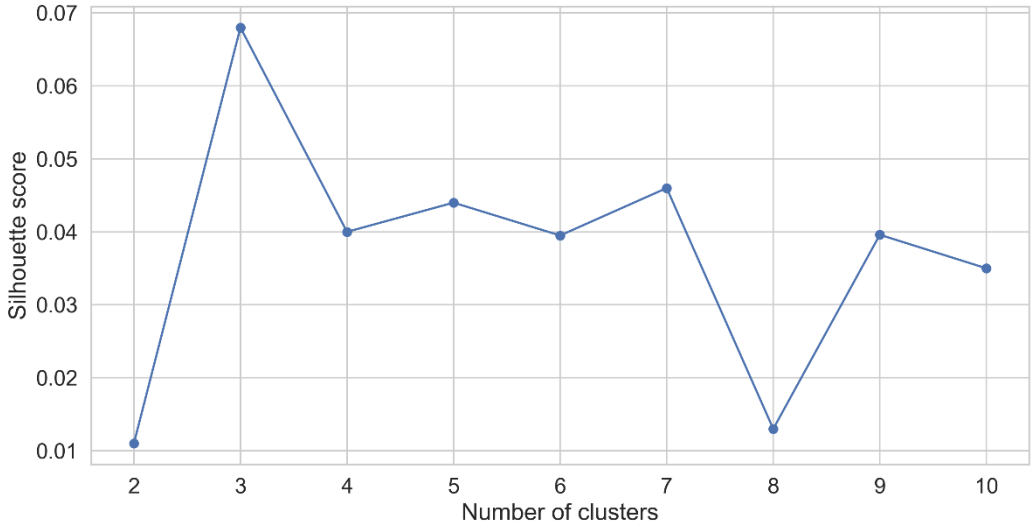

Figure 3. Silhouette Score in the original weight matrix

Analyzing Figure 2, the WCSS minimization function showed a less pronounced decrease in its value from nine clusters onward. Conversely, the Silhouette method (Figure 3), aiming for higher scores (greater cohesion), yielded the best results when *k* equaled 3, 5, and 7. A cluster quantity exceeding five would complicate the analysis, as it would be difficult to find topics or themes that could define precisely what the clusters were about, such as 'politics,' 'macroeconomics,' and 'investments,' among others, so a *k* equal to 3 was tested but produced an unsatisfactory outcome, concentrating almost all instances into a single cluster. Therefore, a decision was made to define a quantity of five groups for further analysis.

It is important to note that in the WCSS analysis (Figure 2), the elbow method does not show a clear point where the curve smooths out, making cluster analysis difficult. The Silhouette method can partially overcome this issue because it is a quantitative and directly comparable measure. However, with the negative linear trend observed in Figure 2 and the low values on the y-axis of Figure 3, it is evident that there is a difficulty in finding well-separated clusters, likely due to the high dimensionality to which the K-means algorithm is subject [Sun et al., 2012].

Subsequently, k and the weight matrix (840 x 612) served as input parameters for the K-means algorithm. Table 3 illustrates the number of tweets allocated to each of the five clusters.

Table 3. Absolute and relative distribution into the five clusters using the original weight matrix

# cluster	Absolute frequency of tweets	Relative frequency of tweets
0	12	1,44%
1	502	59,76%
2	37	4,40%

3	231	27,50%
4	58	6,90%
Total	**840**	**100%**

3.4. Clustering with the reduced matrix through PCA

Before conducting the second clustering, it was necessary to apply PCA to the original weight matrix (840 x 612) and obtain a new matrix with the number of principal components predefined as 2, meaning two columns representing the two orthogonal axes. To determine the best number of clusters for this clustering, the methods WCSS and Silhouette Score were also applied, as shown in Figures 4 and 5.

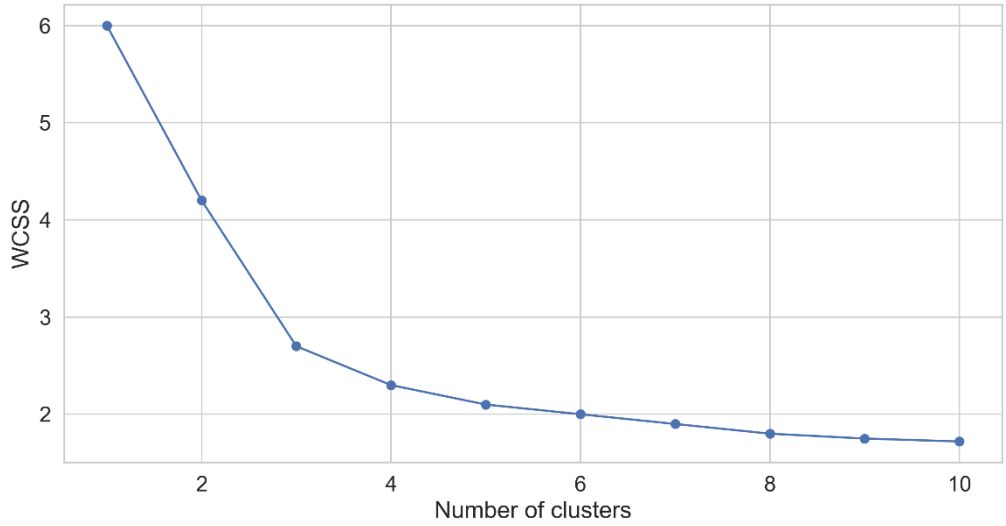

Figure 4. WCSS results in the reduced matrix using PCA

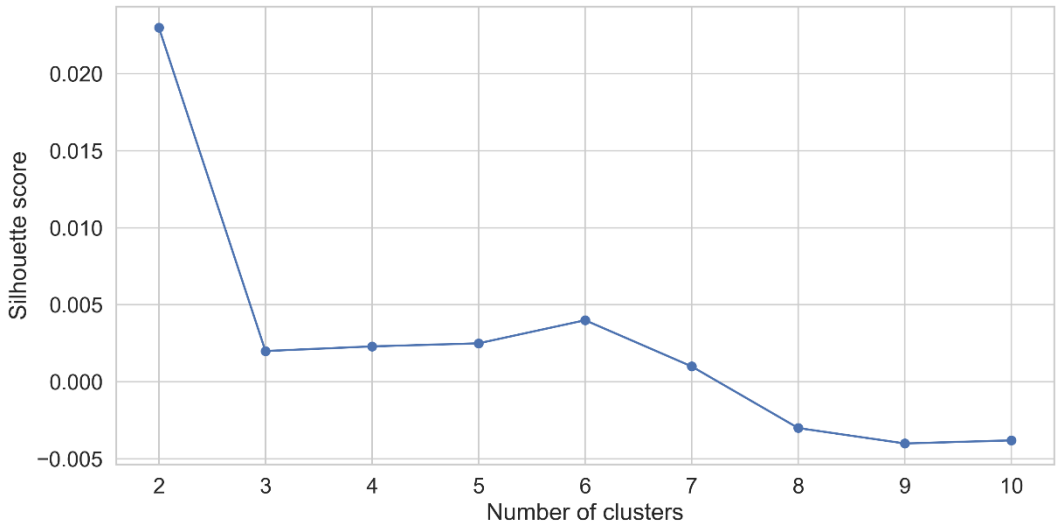

Figure 5. Silhouette score in the reduced matrix using PCA

Although for a *k* value of 2, it yielded good results in both methods, upon conducting the clustering, 98% of the tweets were clustered into a single group, complicating further analyses. Therefore, it was decided that the number of clusters would be set to 3, as the next point showed a more noticeable decline in the minimization method (WCSS value) and stood among the better values in the Silhouette Score.

For Figures 4 and 5, it is noteworthy that the issue of high dimensionality no longer exists, as there are only two components (variables) that primarily aid in the analysis of WCSS. The values from the Silhouette method remain low, as in Figure 3, but there is a notable curve for the WCSS in Figure 4, which did not occur in Figure 2. Thus, it is highlighted again that a number of clusters equal to 3 would be the most appropriate for this case, with a better distribution of tweets among the created groups, avoiding concentration in just one cluster when k = 2.

As input, the K-means algorithm received a *k* value of 3, and the weight matrix was resized by the PCA method (840 x 2). Table 4 displays the arrangement of tweets in each of the groups (0, 1, and 2).

Table 4. Absolute and relative distribution into the three clusters using the reduced weight matrix

# cluster	Absolute frequency of tweets	Relative frequency of tweets
0	581	69,17%
1	240	28,57%
2	19	2,26%
Total	**840**	**100%**

3.5. Comparison of the results

To facilitate comparison among the groups, scatter plots and tables of the most recurring words were generated. For the first clustering, to create a scatter plot, it was necessary to resize the data using PCA and correlate the new resulting matrix with the cluster assigned to each tweet by K-means. On the other hand, for the second clustering, it was only necessary to relate the clusters to the two-dimensional weight matrix initially resized for this method. Figures 6 and 7 exhibit the tweets' positions on the Cartesian plane based on their principal components.

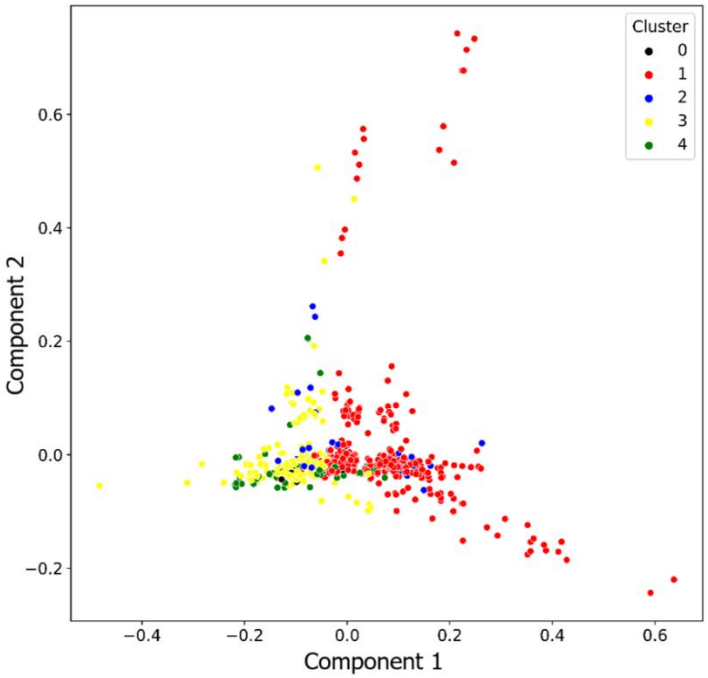

Figure 6. Scatterplot by cluster with the original weight matrix

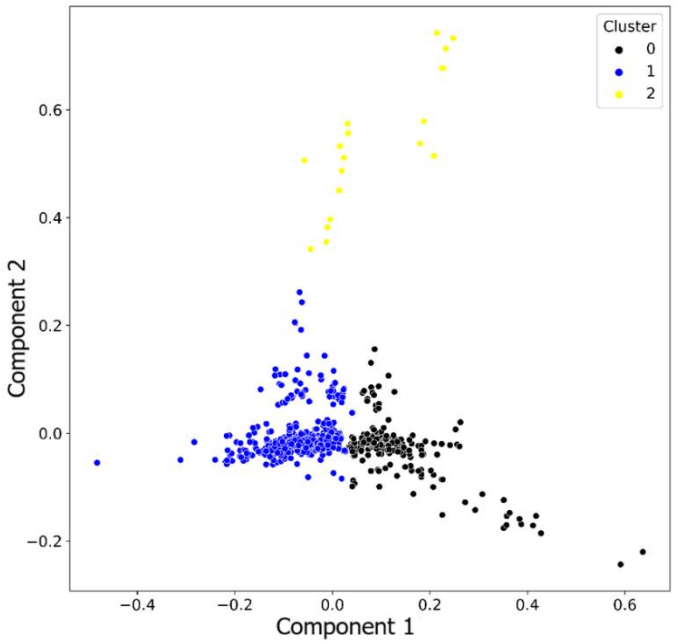

Figure 7. Scatterplot by cluster with the reduced matrix by PCA

As displayed in Figure 6, the clusters with higher data volume, 1 and 3, are spatially positioned in opposite locations, highlighting distinct characteristics. Groups 0, 2, and 4, containing fewer data points and some overlapping points from the predominant groups, are more challenging to identify. This was expected since the first experiment's clustering considered over 600 distinct normalized tokens (variables), resulting in the algorithm's centroids being updated in each iteration within a complex space.

Upon comparing the clusters from Figure 6 to those in Figure 7, the latter resulting from PCA application, a better spatial distinction is noticeable. However, this doesn't necessarily imply that the distributions in the second clustering are superior. Yet, it's notable that when examining both figures, the points of clusters 0 and 2 in Figure 7 mostly correspond to the red points (cluster 1) from the clustering using the original weight matrix. The blue points (cluster 1) in Figure 7 nearly correspond to points from clusters 0, 3, and 4.

Clusters 1 and 3 from the first clustering, representing approximately 87% of the total tweets, were further analyzed. Meanwhile, for the second clustering, the defined groups were 0 and 1. Table 5 presents the most frequent terms for clusters 1 and 3 in the initial clustering strategy, considering the original weight matrix.

Table 5. Frequency distribution of tokens in the total set and each group for the first clustering strategy

Cluster 1				**Cluster 3**			
Token	**Abs. Cluster**	**Rel. Cluster**	**Abs. Data**	**Token**	**Abs. Cluster**	**Rel. Cluster**	**Abs. Data**
Brasil	51	65,38%	78	Presidente	20	54,05%	37
Falir	38	95,00%	40	Governo	19	51,35%	37
Roubo	18	64,29%	28	Preço	18	28,13%	64
Presidente	17	45,95%	37	Bilhões	12	36,36%	33
Dinheiro	17	54,84%	31	PT	11	42,31%	26
Ano	17	48,57%	35	Política	11	42,31%	26
Governo	16	43,24%	37	Roubo	10	35,71%	28
Bilhões	16	48,48%	33	Ano	9	25,71%	35
EUA	15	62,50%	24	Lucro	8	42,11%	19
PT	15	57,69%	26	Quebra	5	33,33%	15
Trabalho	14	63,64%	22				

Legend: Absolute Frequency in the cluster **(Abs. Cluster)**; Relative Frequency in the cluster **(Rel. Cluster)**; Absolute Frequency in the entire dataset **(Abs. Data)**

Upon examining Table 5, although the two clusters exhibit some different words, the most predominant terms in each of them were directly associated with the company, government, and illegalities. Regarding the second clustering strategy,

considering the dataset reduced by PCA, Table 6 displays the distribution results across clusters.

Table 6. Frequency distribution of tokens in the total set and each group for the second clustering strategy

Cluster 0				Cluster 1			
Token	**Abs. Cluster**	**Abs. Data**	**Rel. Cluster**	**Token**	**Abs. Cluster**	**Abs. Data**	**Rel. Cluster**
EUA	15	24	62,50%	Brasil	37	78	47,44%
Estoque	15	15	100,00%	Falir	35	40	87,50%
Brasil	13	78	16,67%	Governo	32	37	86,49%
Poço	12	12	100,00%	Presidente	29	37	78,38%
Combustível	11	29	37,93%	Ano	28	35	80,00%
Hoje	11	31	35,48%	Roubo	27	28	96,43%
Queda	11	14	78,57%	PT	26	26	100,00%
Produção	10	14	71,43%	Dia	26	26	100,00%
Minério	9	16	56,25%	Pagar	25	27	92,59%
Gasolina	8	30	26,67%	Dinheiro	23	31	74,19%
EUA	15	24	62,50%	Bilhões	21	33	63,64%

Legend: Absolute Frequency in the cluster **(Abs. Cluster)**; Relative Frequency in the cluster **(Rel. Cluster)**; Absolute Frequency in the entire dataset **(Abs. Data)**

However, it is noted that Table 6 identified in the second experiment that Cluster 0 is more related to the marketing and production of oil, with a more neutral tone, while in Cluster 1, recurring words are associated with political-governmental issues, highlighting a greater polarization among the profiles that posted, with negative sentiments being prominent.

In summary, it is evident that the second clustering, after applying PCA to the weight matrix obtained from TF-IDF, makes word frequency analysis more apparent, achieving a clearer separation among the content of the posts and providing some insight into the profiles of individuals who posted, shared, or retweeted.

Thus, it is understood that reducing the data matrix with PCA significantly enhanced the interpretation of the results, both quantitatively (in terms of cluster investigation metrics) and qualitatively. However, there are several challenges for clustering techniques applied to the financial or business market. For instance, Gupta et al. (2020) assert that text mining still faces some hurdles, such as restricted access to confidential data, sarcastic and informal language, highly unstructured and redundant data, sarcasm, vernacular language, lack of well-defined financial lexicon lists, absence of dynamic text analysis models, and the need to cluster results across domains.

In the routine of the investor, this has a significant impact, considering a human can't read and categorize the same amount of text so rapidly. Through algorithms, investors can gain advantages by assessing the most frequently discussed topics in a text set and mapping their potential implications in the financial market. Such implications

may signal highs or lows in the stock price of a company, indicating, for instance, the best time to buy or sell assets of that company, leading to potential gains surpassing the Ibovespa index. As Derakhshan and Beigy (2019) affirm, stock prices are affected by various factors, including macroeconomics and diverse news, and the expansion of the internet and social networks is being monitored to track people's opinions and emotions about a company or stock.

In this same context, it has been established that sentiment analysis and attention variables have good predictive power for stock volatility, especially concerning texts from investors, including the number of posts made by them on social network platforms [Audrino et al., 2020].

Conclusions

The correlation between opinions and news shared in tweets and the tracking of profiles through clustering techniques constitutes a growing field of study. The possibilities have expanded with the advent of new artificial intelligence platforms that facilitate text preprocessing or even intelligent generation. This study served as a starting point for comprehensively analyzing the state-owned company Petrobras, assessing potential tweet characteristics for grouping similar profiles or content.

This work revealed that the sentiments and news expressed in social media can be separated using unsupervised machine learning techniques, aiding in investment decisions and the perception of investors and society regarding a market or a specific company. Also, the influence of politically charged texts and polarization can be critical in the analysis of companies and groups of people formed by clustering techniques. Political polarization might impact public perception of certain companies, sectors, or government policies, leading to market volatility and affecting the value or credibility of companies, especially state-owned enterprises.

The research's limitations lie in the direct qualitative analysis of the most frequent terms, lacking more sophisticated statistical techniques. It's also recognized that a more in-depth investigation into the clusters formed by sentiment analysis using established libraries or frameworks in the literature, such as the RoBERTa model proposed by the Hugging Face community, is feasible.

7. References

Adegboye, A., and Kampouridis, M. (2021). Machine learning classification and regression models for predicting directional changes trend reversal in FX markets. “Expert Systems with Applications”, 173, 114645. https://doi.org/10.1016/j.eswa.2021.114645

Akita, M., and da Silva, E. J. (2023). “Desenvolvimento de modelo para predição de cotações de ação baseada em análise de sentimentos de tweets”. Anais Do 1º Seminário de Ciência de Dados Do IFSP, 51–58.

Al-jabery, K. K., Obafemi-Ajayi, T., Olbricht, G. R., and Wunsch Ii, D. C. (2020). Evaluation of cluster validation metrics. *In* Computational Learning Approaches to Data Analytics in Biomedical Applications (p. 189–208). Elsevier. https://doi.org/10.1016/B978-0-12-814482-4.00007-3

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., and Perona, I. (2013). An extensive comparative study of cluster validity indices. “Pattern Recognition”, 46(1), 243–256. https://doi.org/10.1016/j.patcog.2012.07.021

Audrino, F., Sigrist, F., and Ballinari, D. (2020). The impact of sentiment and attention measures on stock market volatility. “International Journal of Forecasting”, 36(2), 334–357. https://doi.org/10.1016/j.ijforecast.2019.05.010

B3. (2020, December 14). “B3 divulga estudo sobre os 2 milhões de investidores que entraram na bolsa entre 2019 e 2020”. Http://Www.B3.Com.Br/Pt_br/Noticias/Investidores.Htm.

Braga, F. dos R. (2018). Extração semiautomática de taxonomia para domínios especializados usando técnicas de mineração de textos. “Ciência Da Informação”, 45(3), 175–186. https://doi.org/10.18225/ci.inf.v45i3.4056

Carosia, A. E. O., Coelho, G. P., and Coelho, G. P. (2020). Analyzing the Brazilian Financial Market through Portuguese Sentiment Analysis in Social Media. “Applied Artificial Intelligence”, 34(1), 1–19. https://doi.org/https://doi.org/10.1080/08839514.2019.1673037

Chen, J. M., Rehman, M. U., and Vo, X. V. (2021). Clustering commodity markets in space and time: Clarifying returns, volatility, and trading regimes through unsupervised machine learning. “Resources Policy”, 73, 102162. https://doi.org/https://doi.org/10.1016/j.resourpol.2021.102162

da Silva, E. S., Almagro, H. U., Queiroz, S. S., Henrique, M. R., and Soares, W. A. (2020). Estudo comparativo entre a rentabilidade de seis empresas no mercado de ações pela escola fundamentalista e técnica. “Revista OIDLES”, 14(28), 112–142. http://hdl.handle.net/20.500.11763/oidles28escola-fundamentalista-tecnica

de Oliveira, F. A., Nobre, C. N., and Zárate, L. E. (2013). Applying Artificial Neural Networks to prediction of stock price and improvement of the directional prediction index – Case study of PETR4, Petrobras, Brazil. “Expert Systems with Applications”, 40(18), 7596–7606. https://doi.org/10.1016/j.eswa.2013.06.071

Derakhshan, A., and Beigy, H. (2019). Sentiment analysis on stock social media for stock price movement prediction. “Engineering Applications of Artificial Intelligence”, 85, 569–578. https://doi.org/10.1016/j.engappai.2019.07.002

Gupta, A., Dengre, V., Kheruwala, H. A., & Shah, M. (2020). Comprehensive review of text-mining applications in finance. “Financial Innovation”, 6(1), 39. https://doi.org/10.1186/s40854-020-00205-1

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction R. “Expert Systems with Applications”, 124, 226–251. https://doi.org/10.1016/j.eswa.2019.01.012

Info Money. (2017, January 24). “As 30 contas no Twitter que todo investidor deve seguir em 2017”. Https://Www.Infomoney.Com.Br/Mercados/as-30-Contas-No-Twitter-Que-Todo-Investidor-Deve-Seguir-Em-2017/.

Info Money. (2022, January 11). “B3 (B3SA3) tem volume financeiro negociado no mercado à vista recorde em 2021”. Https://Www.Infomoney.Com.Br/Mercados/B3-B3sa3-Tem-Volume-Financeiro-Negociado-No-Mercado-a-Vista-Recorde-Em-

2021/. https://www.infomoney.com.br/mercados/b3-b3sa3-tem-volume-financeiro-negociado-no-mercado-a-vista-recorde-em-2021/

James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J. (2023). An Introduction to Statistical Learning with Applications in Python (Vol. 1). https://www.statlearning.com/

Jesus Júnior, L. B. de, Sarti, F., and Ferreira Júnior, H. de M. (2017). Petrobras, política de conteúdo local e maximização de valor para o acionista: uma sugestão de interpretação. "Economia e Sociedade", 26(2), 369–400. https://doi.org/10.1590/1982-3533.2017v26n2art4

Katayama, D., and Tsuda, K. (2020). A Method of Using News Sentiment for Stock Investment Strategy. "Procedia Computer Science", 176, 1971–1980. https://doi.org/https://doi.org/10.1016/j.procs.2020.09.333

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Ying Wah, T., and Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. "Expert Systems with Applications", 41(16), 7653–7670. https://doi.org/10.1016/j.eswa.2014.06.009

Kononova, O., He, T., Huo, H., Trewartha, A., Olivetti, E. A., and Ceder, G. (2021). Opportunities and challenges of text mining in materials research. "IScience", 24(3), 102155. https://doi.org/10.1016/j.isci.2021.102155

Kumbure, M. M., Lohrmann, C., Luukka, P., and Porras, J. (2022). Machine learning techniques and data for stock market forecasting: A literature review. "Expert Systems with Applications", 197(December 2021), 116659. https://doi.org/10.1016/j.eswa.2022.116659

Lima, M. L., Nascimento, T. P., Labidi, S., Timbó, N. S., Batista, M. V. L., Neto, G. N., Costa, E. A. M., and Sousa, S. R. S. (2016). Using Sentiment Analysis for Stock Exchange Prediction. "International Journal of Artificial Intelligence & Applications", 7(6), 59–66.

Mohamed, A. A. (2020). An effective dimension reduction algorithm for clustering Arabic text. "Egyptian Informatics Journal", 21(1), 1–5. https://doi.org/10.1016/j.eij.2019.05.002

Money Times. (2021, January 11). "Volume movimentado pela B3 salta 71%, em 2020, e quase empata com o PIB pela primeira vez." Https://Www.Moneytimes.Com.Br/Volume-Movimentado-Pela-B3-Salta-71-Em-2020-e-Quase-Empata-Com-o-Pib-Pela-Primeira-Vez/.

Naghizadeh, A., and Metaxas, D. N. (2020). Condensed Silhouette: An Optimized Filtering Process for Cluster Selection in K-Means. "Procedia Computer Science", 176, 205–214. https://doi.org/10.1016/j.procs.2020.08.022

Nam, K., and Seong, N. (2019). Financial news-based stock movement prediction using causality analysis of influence in the Korean stock market. "Decision Support Systems", 117, 100–112. https://doi.org/https://doi.org/10.1016/j.dss.2018.11.004

Nizer, P. S. M., and Nievola, J. C. (2012). Predicting published news effect in the Brazilian stock market. "Expert Systems with Applications", 39(12), 10674–10680. https://doi.org/https://doi.org/10.1016/j.eswa.2012.02.162

Oliveira, N., Cortez, P., and Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. “Expert Systems with Applications”, 73, 125–144. https://doi.org/https://doi.org/10.1016/j.eswa.2016.12.036

Pompella, M., and Dicanio, A. (2017). Ratings based Inference and Credit Risk: Detecting likely-to-fail Banks with the PC-Mahalanobis Method. “Economic Modelling”, 67, 34–44. https://doi.org/10.1016/j.econmod.2016.08.023

Raschka, S. (2015). Python Machine Learning (1st ed.). Packt Publishing Ltd.

Sarram, G., and Ivey, S. S. (2022). Evaluating the potential of online review data for augmenting traditional transportation planning performance management. “Journal of Urban Management”, 11(1), 123–136. https://doi.org/https://doi.org/10.1016/j.jum.2022.01.001

Seong, N., and Nam, K. (2021). Predicting stock movements based on financial news with segmentation. “Expert Systems with Applications”, 164, 113988. https://doi.org/10.1016/j.eswa.2020.113988

Shi, Y., Tang, Y., and Long, W. (2019). Sentiment contagion analysis of interacting investors: Evidence from China’s stock forum. “Physica A: Statistical Mechanics and Its Applications”, 523, 246–259. https://doi.org/https://doi.org/10.1016/j.physa.2019.02.025

Singh, R., and Srivastava, S. (2017). Stock prediction using deep learning. “Multimedia Tools and Applications”, 76(18), 18569–18584. https://doi.org/10.1007/s11042-016-4159-7

Sun, W., Wang, J., and Fang, Y. (2012). Regularized k-means clustering of high-dimensional data and its asymptotic consistency. “Electronic Journal of Statistics”, 6, 148-167. https://doi.org/10.1214/12-EJS668

Sun, Y., Liu, X., Chen, G., Hao, Y., and Zhang, Z. (Justin). (2020). How mood affects the stock market: Empirical evidence from microblogs. “Information & Management”, 57(5), 103181. https://doi.org/https://doi.org/10.1016/j.im.2019.103181

Twitter Inc. (n.d.). Search Tweets: standard v1.1. Retrieved December 28, 2021, from https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets

Ulfenborg, B., Karlsson, A., Riveiro, M., Andersson, C. X., Sartipy, P., and Synnergren, J. (2021). Multi-assignment clustering: Machine learning from a biological perspective. “Journal of Biotechnology”, 326, 1–10. https://doi.org/10.1016/j.jbiotec.2020.12.002

Yadav, A., Jha, C. K., Sharan, A., and Vaish, V. (2019). “Sentiment Analysis of Financial News Using Unsupervised and Supervised Approach”, In: *International Conference on Pattern Recognition and Machine Intelligence* (pp. 311–319). https://doi.org/10.1007/978-3-030-34872-4_35